Using Speech-to-Text Applications for Assessing English Language Learners' Pronunciation: A Comparison with Human Raters

Akiyo Hirai and Angelina Kovalyova

Abstract With the growing influence of technology in the English as Foreign Language (EFL) classroom, automatic speech recognition (ASR) has been receiving a great deal of attention as a tool for pronunciation practice. In particular, the immediate feedback it provides about the level of accentedness and comprehensibility of a user's speech keeps the interest growing. This chapter focuses on the use of speech-to-text (STT) applications, a variation of ASR technology, to explore the potential of using such applications to evaluate the pronunciation of adult EFL learners with different first languages (L1). The chapter discusses the use of ASR in the English language classroom context. It focuses on the accuracy and reliability of five current STT applications (Google Docs' Voice Typing, Windows 10 Dictation, Apple's Dictation, a website service "Dictation.io," and the iOS application "Transcribe"). The chapter concludes that, with a 50–70% accuracy rate, speech recognition still has room for improvement when used by EFL learners. However, it is the absence of perfect speech recognition that helps EFL learners identify their pronunciation errors. Even more so, teachers can rely on STT applications as the pronunciation assessment of these applications was found to be consistent with the pronunciation evaluation by human raters.

Keywords Automatic speech recognition (ASR) · Speech-to-text applications · Pronunciation · Accuracy · Adult EFL learners

A. Hirai (✉) · A. Kovalyova
University of Tsukuba, Tsukuba, Japan
e-mail: hirai.akiyo.ft@u.tsukuba.ac.jp

M.-M. Suárez, W. M. El-Henawy (eds.), *Optimizing Online English Language Learning and Teaching*, English Language Education 31,
https://doi.org/10.1007/978-3-031-27825-9_17

1 Background of ASR and STT Technology

The COVID-19 pandemic has left a definitive mark on how humans interact with technology. Rapid digitalization of many spheres of life has created a new normal. It has forced many industries, including education, to test new digital solutions to keep up with increasing demand. One such growing area of interest concerns integrating automatic speech recognition (ASR) technology into our lives. This has seen significant improvements since the early 90s, when decoding the human voice using a computer was seen as experimental (Kincaid, 2018).

Traditional ASR technology involves a process whereby human speech is received, decoded, and transformed into text by a computer program as a part of human-computer interaction (Microsoft, 2004). ASR technology has played an active role in many areas of our lives, beyond our expectations. Digital assistants such as Apple's Siri and Amazon's Alexa help us navigate our smart homes and use digital services. Dictation and speech-to-text (STT) tools assist us in our work environments, allowing us to quickly write down meeting minutes or important ideas using only our voice. Voice calls to public or private services are often accompanied by computer-assisted dialogue that requires callers to answer questions to verify their identity or make an appointment (REV.com, 2020). Moreover, recent advancements in artificial intelligence (AI) and natural language processing (NLP) have pushed ASR technology to new limits, bringing hope that language recognition will become a ubiquitous service.

As the quality of digital services grows, it is natural that ASR technology has shown great potential in the context of education, STT technology being one of its most common applications. STT technology—sometimes referred to as speech-to-text recognition (STR) technology—is an extension of ASR technology in that human speech recognized by ASR software can be transcribed into text in real-time (Hwang et al., 2012, p. 368). In other words, text spoken by a person is displayed in a word processor or other text applications, allowing us to see how accurately the ASR technology recognizes human speech. Some common examples of STT applications include Google Docs' Voice Typing function, Apple's Dictation tool, Dragon Speech Recognition Solutions, and Speechnotes.

Even though these applications are not necessarily designed for language learning purposes, they are becoming a prominent tool in language learning. Multiple studies have shown that ASR tools can support classroom activities and have great potential to assist with language learning (e.g., Hwang et al., 2012; Liakin et al., 2014). Therefore, the conceptual framework of this chapter will focus on observing the potential of STT applications in recognizing the speech of adult non-native speakers (NNS) of English with various first language (L1) backgrounds. The study, used as a backbone for this chapter, illustrates the correlation between human and machine evaluation of NNS speech and discusses the accuracy and reliability of the five STT applications, providing practical advice for their classroom use.

2 STT Technology in EFL Classroom

References to educational research related to ASR technology date back to Coniam (1998), who explored the voice recognition ability of Dragon Systems software by conducting an experiment whereby a group of English language learners read a text to the ASR program. The paper concluded that speech recognition at that time still needed to be trained by every single speaker to be effective. Since then, several research attempts have been made to test various aspects of ASR technology. In general, earlier studies focused on observing correlations between human pronunciation scores and ASR software evaluation and analyzing whether human speech could be successfully assessed at all. In contrast, more recent studies already aim at understanding how ASR software detects pronunciation errors and how similar this process is to human assessment (O'Brien et al., 2018). In the foreign language learning context, the focus is largely on finding ways to help language learners improve their pronunciation

Liakin et al. (2014), for example, focused on helping learners of French improve the pronunciation of the French /y/ sound by using Nuance Dragon Dictation's ASR technology. The experiment involved three groups of learners of French: the non-ASR group, which practiced pronunciation while receiving feedback from the teacher; the ASR group, which practiced pronunciation and received written feedback from an ASR application; and the control group, which practiced pronunciation with a teacher and received no feedback. The group that practiced pronunciation with ASR weekly and received written feedback was the only group that significantly improved pronunciation of the French /y/ sound.

Besides helping to improve individual pronunciation, ASR technology can be used to encourage students' interactive speaking practice. Ahn and Lee (2016) utilized *English 60 Junior,* a specially designed mobile-based learning system with integrated Google Voice ASR, to allow a group of Korean middle-school students to practice English conversation. Students noted that written feedback provided by the ASR technology became a valuable tool for analyzing their pronunciation and that the application gave them more opportunities to practice speaking and made doing so more interactive (pp. 783–784).

Evers and Chen (2020) observed how different learning styles (visual/verbal) and other types of feedback affected English as a Foreign Langauge (EFL) adult learners who practiced pronunciation using ASR technology. Out of three groups (1–receiving pronunciation feedback from a teacher, 2–receiving feedback from ASR and peers, and 3–receiving feedback only from ASR), the second group showed the most significant improvement in pronunciation performance in both learning styles (conversation/verbal and reading/visual). McCrocklin (2019) also compared the pronunciation performance of different L2 EL groups (one with entirely face-to-face instruction and another with hybrid instruction where half the time was devoted to using the ASR dictation program (Windows Speech Recognition). While the results didn't show any statistically significant differences

between the groups, the study concluded that ASR technology could complement face-to-face pronunciation training.

Thus, ASR technology is gradually being utilized as a supplementary tool in language learning, helping with accent training and providing additional opportunities for speaking practice.

3 Constraints of STT Technology

From a technical standpoint, the claim of high speech recognition accuracy is the most significant selling point for ASR tools' progress. In 2017, Google claimed to have reached a 95% accuracy rate for U.S. English (Worthy, 2019), while Microsoft achieved 93.1% accuracy (Hachman, 2017). These numbers suggest that ASR technology has already reached human-like comprehension and can recognize human speech with minimal errors.

However, it is important to acknowledge the experimental conditions in which such high accuracy rates were achieved. Gevirtz (2019) pointed out that both Google and Microsoft trained and validated their ASR systems using the National Switchboard Corpus (Godfrey & Holliman, 1993), a small database of phone calls in U.S. English carefully prepared for linguistic research. Such a data set is somewhat limiting as it does not include today's many English language varieties. Thus, judging from the results, ASR systems cannot offer the same potential to the wider English-speaking community, let alone to learners of English, whose language still needs improvement.

When it comes to accent recognition by STT tools, some variation in accuracy rate has been reported even among native English speakers (Koenecke et al., 2020, pp. 7684–7685). It is thus conceivable that non-native speech would heavily affect ASR performance. A study commissioned by the Washington Post (Harwell, 2018) discovered that speech performance in non-native accents (i.e., Spanish, Chinese, and Indian) significantly affects the accuracy rate of English speech recognition. Google Assistant and Amazon's Alexa performance was up to 30% less accurate when non-American accents were used with their speech recognition systems compared to native speakers, which had 91.8% and 91% accuracy rates, respectively. It has become clear that modern ASR systems need to expand their data sets in order to accommodate a wider population.

Besides the ASR's accuracy rate variation among native speakers (NS) or between NS and NNS, its accuracy rate also varies significantly depending on other factors. The most common issues hindering ASR's accuracy include different kinds of background noise, the use of rare words and jargon, multiple speakers, non-fluency features, and lack of training to have ASR systems get used to recognizing the user's pronunciation (Gevirtz, 2019; Ito, 2014; Jarnow, 2016; Way et al., 2008). Thus, in a situation whereby two or three people with accented English are having a meeting or discussion and are using industry-specific jargon, ASR would provide

a transcription of the conversation with a significant number of errors that would require further review and correction. The National Institute of Informatics in Japan points out that raw human interaction is too chaotic for speech-recognition systems that can provide around a 90% accuracy rate only when text designed for speech recognition has been prepared ahead of time (Ito, 2014, p. 10). This is because a speaker, when producing spontaneous speech, can suffer from non-fluency features such as false starts, hesitations, and repetition that can leave the speech disorganized and difficult to analyze.

These limitations of ASR technology may leave teachers discouraged about the success of speech recognition systems' use in a classroom. Also, the low accuracy rate of speech transcriptions due to frequent grammatical and lexical errors made by language learners will, in turn, leave students demotivated. For example, Bajorek (2017), in her review of modern software for practicing pronunciation, attempted to analyze the pronunciation presentations of Rosetta Stone, Duolingo, Babbel, and Mango Languages. As a result, Bajorek concluded that, despite their potential, each application has specific limitations; thus, they do not give enough support for pronunciation training for a language learner using these applications independently. Therefore, she commented that it was no surprise that teachers are unaware of how to use ASR technology effectively or are hesitant about how they should use it (Bajorek, 2018, p. 3).

Considering Bajorek's remark that ASR technology is still developing and that not all tools would be useful, McCrocklin (2015, p. 131) also recognized the current limitations of ASR technology, in that such tools may not give perfect feedback because some are too sensitive to pronunciation errors while others are too forgiving. This is where students can benefit from the guided feedback and support provided by a teacher. In addition, when using STT technology in an EFL setting, it is important to be aware of the benefits and limitations of speech recognition technology.

4 A Study on Adult NNS Speech Recognition: Current Experiment Research Findings

ASR software holds undeniable potential for language-learning purposes. Still, since the success of its execution heavily depends on the software's capabilities and the surroundings, it is essential to understand the degree to which such software can replace human feedback. If we aim to use ASR technology in an EFL classroom, we cannot blindly rely on reported accuracy rates as current ASR software is assessed through the evaluation of speech by NS (Gevirtz, 2019). Thus, it is important to explore the speech recognition context of speech by NNS and understand how accurate the existing ASR technology can be and how reliable it is compared with human speech assessment.

4.1 Applications Analyzed in the Study

These days, there are many different ASR options, from free, easy-to-use applications to commercial tools for professionals that focus on specific jargon. How can a teacher know which tool would be safe to use in a classroom?

Our study focused on five STT applications available to many users. These applications include Google Docs' Voice Typing, Windows 10's Dictation, Apple's Dictation, the Dictation.io web service, and the "Transcribe" iOS application. Each of these tools is free (although "Transcribe" has an additional subscription option), and they cross various platforms so that users do not feel restricted in their choices. They are described based on their performance as of spring 2020.

A brief overview of these applications can provide information about the capabilities and limitations of current, readily available ASR technology. If a teacher is familiar with Google's services, trying out the Voice Typing function of Google Docs might be the easiest option since it is a feature of the cloud-based Google Docs and Google Slides services (accessed through a Chrome browser). One needs to open a new Google Docs file, select the "Tools" tab, click on "Voice Typing," and start speaking; the program will then begin to write down the user's utterances immediately. Another option is Apple's Dictation, a built-in ASR tool available on iOS devices. This can be accessed through the "Dictation" settings on a laptop or by tapping on the microphone sign on an iPhone or an iPad keyboard. Likewise, Windows 10's Dictation is part of the Windows software package and can be accessed through the "Speech" section of the platform's settings. It is necessary to follow up by pressing a combination of the Windows logo key and "H" to prompt the dictation service. Next, the Dictation.io web service can be accessed through a browser, regardless of the operating system or browser type, and perhaps provides the easiest interface and most user-friendly experience. These four STT tools offer synchronous speech transcription through which a person can speak and instantly see a transcription of what they have said. The final of the five analyzed STT tools, an iOS mobile application, "Transcribe," is an example of asynchronous ASR analysis. A user has to upload an audio file into the application to receive an analysis and a transcription of the speech. The application also predicts transcription accuracy in percent (%).

Each of these applications has its strengths and weaknesses. They all support many languages; for example, Google Docs' Voice Typing supports up to 119 language varieties (Google., n.d.), and Dictation.io attempts transcription in 134 language varieties. This is especially helpful in an EFL classroom, for example, when a student speaks a certain variety of English (or even wants to train a particular accent) as these services can differentiate between the pronunciation of Canadian or New Zealand English, English in the Philippines, and so forth. These STT applications also generally require an internet connection to provide a quality service as they use cloud storage to increase computational power for ASR analysis (Altviz. co., 2019, p. 4). Finally, it is worth remembering that some STT applications (i.e., Windows 10's Dictation) can be trained to better recognize an individual's

pronunciation over time. This might have its benefits and challenges too. When the application becomes accustomed to a user's pronunciation, it may give learners correct transcriptions even though their pronunciation remains accented. Overall, when using any of these applications, a teacher must first test it to evaluate whether and how STT technology could be incorporated into pronunciation practice in their EFL classroom.

4.2 Accuracy of Pronunciation Assessment with STT Applications

To estimate ASR accuracy from adult NNS English speech, we conducted a research project to test the five STT applications. Thirty university students, all NNS of English (18 Japanese, 4 Chinese, 3 Korean, and 5 other nationalities including Czech, Hungarian, Pakistani, French, and Taiwanese) were asked to respond to four test tasks, having their speech first recorded and then transcribed by each of the five applications, as well as by a human. Regarding students' English proficiency, at the beginning of the experiment, the participants were asked about their language learning background, including whether they had taken a language proficiency test. According to the questionnaire, 4 learners had attained the A1 level of English proficiency (CEFR framework), 4 attained B1, 15 attained B2, and 7 attained C1 level. None of the participants reported A2 or C2 levels of English proficiency.

The tasks involved reading out loud short sentences with around 25 words each (Task 1: ReadS), reading out loud a long passage with approximately 100 words (Task 2: ReadL), retelling a long passage that had previously been read (Task 3: Retell), and answering three questions (Task 4: QA). Tasks 1 and 4 also included loan words from Japanese, such as "haiku," "bukatsu," "sempai," and "kouhai." These tasks were designed to test different aspects of ASR transcription ability in an NNS speech.

As explained earlier, the five STT applications involved were the Voice Typing function of Google Docs, Windows 10's Dictation tool, Apple's Dictation, the Dictation.io web service, and the "Transcribe" iOS application. These were chosen for their accessibility and variety. The speech was recorded in a quiet room using an iPhone Voice Memo application and a microphone.

4.2.1 Effect of Speech Task on Transcription Accuracy

The accuracy rate of each application was determined by calculating the number of correctly transcribed words by each STT application against the total number of words received from human-made transcriptions. As shown in Table 1, the results varied depending on the application and the type of speech task performed, with an average accuracy rate of 50–70% across the five STT applications. Of these,

Table 1 Transcription accuracy rates of nonnative speech using five STT applications (N = 30)

	Task 1: ReadS		Task 2: ReadL		Task 3: Retell		Task 4: QA		Total	
Application	*M* (%)	*SD*	*M* (%)	*SD*	*M* (%)	*SD*	*M* (%)	*SD*	*M* (%)	*SD*
Google	64.46	19.90	64.28	19.53	57.76	22.77	65.85	17.84	**63.09**	19.76
Apple	45.38	18.38	52.44	17.45	44.14	22.50	53.94	16.83	**48.97**	18.87
Windows 10	60.97	17.59	69.75	13.87	66.58	20.13	70.54	13.94	**66.96**	16.55
Dictation	58.42	19.41	57.39	19.23	50.74	24.11	60.33	16.50	**56.72**	19.76
Transcribe	65.97	18.16	68.11	16.72	68.80	15.97	71.38	12.64	**68.57**	15.66
Total	**59.04**	19.58	**62.39**	18.19	**57.60**	22.68	**64.41**	16.57		

ReadS (reading short sentences), ReadL (reading a passage), Retell (retelling the passage), and QA (answering questions)

"Transcribe," and Windows 10's Dictation tool showed the best performance (68.57% and 66.96%, respectively); this was nearly 20% higher than Apple's Dictation (48.97%). Thus, there was a variation in transcription accuracy across the STT applications. In addition, when transcribing the speech of NNS in English, the accuracy was significantly lower than the industry average for NS speech in English, as mentioned in section 3.

From the viewpoint of the type of speech (i.e., tasks), on average, the Retell task showed the lowest accuracy result (57.6%). This can be explained by the fact that, compared with the other three tasks, the Retell task had a higher chance of being affected by various intrinsic aspects of natural spontaneous speech, such as self-correction, repetition, hesitation, and false starts. In other words, students had to produce a long string of speech as the content and amount to be retold had been specified in the original passage. In addition, the ReadS task (reading short sentences) was relatively poorly transcribed (59.04%), perhaps because each sentence was too short for the STT applications to predict the following words. Also, the sentences contained loan words, which the applications might not transcribe correctly in their English mode.

An additional analysis of the interaction between speech task type and STT application was carried out using a two-way repeated-measures ANOVA test to see if different types of speech production would influence the ASR process. The result revealed a significant interaction between tasks and applications (F (7.36, 213.56) = 4.42, $p < .001$, $\eta_p^2 = 0.13$), meaning that the type of speaking task (i.e., various aspects of speech) affected the accuracy rate of transcription of each STT application differently. With further analysis of multiple comparisons, it was found that Windows 10's Dictation tool showed better results when transcribing the ReadL tasks (69.75%) containing error-free, syntactically coherent sentences (see Fig. 1). However, it was relatively weak in ReadS tasks (60.97%), indicating that Windows 10 can increase the prediction of words used next in long strings of natural speech, but it may have difficulty doing so in such short strings. On the other hand, the performance of "Transcribe" was relatively stable across the different tasks and was the best performing of the five applications.

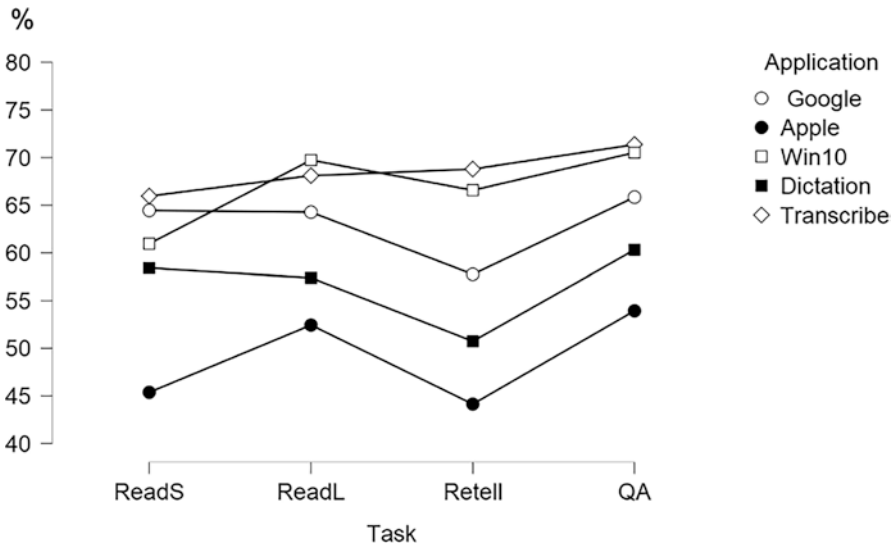

Fig. 1 Comparison of five applications across four speech tasks. (ReadS (reading short sentences), ReadL (reading a passage), Retell (retelling the passage), and QA (answering questions))

These results help us understand that STT applications have strengths and weaknesses depending on the type of speech. A heavier number of transcription errors appears from using a lexicon that does not belong to the language model of ASR for a particular language (such as saying foreign words when using an ASR system for English). Transcription errors are also significant when the flow of speech is interrupted by repairs, repetition, hesitation, and other non-fluency features common in spontaneous speech. Having avoided these issues, transcription performance still suffers from pronunciation errors, which many NNS make in even prepared or simple read-aloud speech.

Each of these issues is inherent in English language learners to a different degree, depending on whether they use a word from their L1 due to a lack of English vocabulary knowledge, make repairs as they try to correct their grammar, or have a stronger accent or come from a culture where their L1 phonetic alphabet drastically differs from English.

4.2.2 Effect of Pronunciation Features on Transcription Accuracy

Despite many factors that affect the accuracy of STT transcriptions, transcription errors can help EFL learners assess their pronunciation, especially when they produce spontaneous speech. During this research project, it was found that automatic transcription was affected by particular pronunciation features of NNS. For example, Japanese speakers, while performing speaking tasks, maintained some

pronunciation errors attributed to the Japanese phonetic alphabet (L1) in English (L2) (Koon, 2018; Vaughn et al., 2018).

One of the more obvious errors has its roots in the influence of *katakana*, which is a set of Japanese words adopted from other languages and made to sound somewhat like the original word from another language (a typical example of this is the word "アイスクリーム" borrowed from the English "ice cream" and pronounced as /ʌisukur'i:mu/). The issue lies in the transfer of *katakana* Japanese pronunciation into English pronunciation, which happens by attaching extra vowels after every consonant. In this way, in our experiment, the word "etiquette" (/ˈɛtɪkɛt/) became /ɛtʃikɛtto/ or "bank" (/bˈæŋk/) became /bʌnku/. Accordingly, an ASR system for the English language offered an alternative English word as a transcription that matched the pronounced form. For example, the pronunciation of /ɛtʃikɛtto/ returned "educate" or "adequate."

Problematic pronunciation errors become especially clear with pronunciation errors in minimal pairs—words that differ only in one phonological element, such as "fan" and "van." Similarly, considering the difficulty of distinguishing /l/ from /r/ in Japanese, it was no surprise that STT applications often mistranslated words containing those phonemes when pronounced by Japanese EFL learners. Transcription errors like this can provide valuable feedback for EFL learners with various L1 backgrounds to locate pronunciation errors as the wrongly transcribed words will point to the deviation from the pronunciation norm (i.e., more recognized pronunciation varieties) (Table 2).

4.3 Reliability of Pronunciation Assessment with STT Applications

Considering the above, it is worth examining whether STT applications can assess the pronunciation of English by NNS. Specifically, does an STT application's pronunciation assessment correlate with that of a human rater? How much can we trust the technology? To answer these questions, a human rater also evaluated the

Table 2 Examples of transcription errors in words with /l/ and /r/ sounds

Original	Wrong transcriptions
Lunch	Branch, ranch
Play	Pray
School	Scooter
Reading	Leading
Sleepy	Sweet
Culture	Carter
Remember	New member

participants' speech. Then the evaluation scores were compared with the accuracy scores of the STT applications.

Regarding the assessment strategies of STT applications, it has been noted that these are rather sensitive to spontaneous speech and stronger accents (Gevirtz, 2019). Thus, a pronunciation assessment rubric for a human rater was first created with the same issues in mind, focusing on the frequency of pronunciation errors, prosodic features, and accent strength. The rubric contained a 4-point scale, where 4 was the highest score (Table 3).

Evaluation of pronunciation by a human rater was conducted by assigning a pronunciation score to each participant's performance in the ReadS, Retell, and QA tasks. ReadL (Task 2) was excluded because the type of task was considered similar to Task 1 in this assessment context. Approximately one-quarter of the participants' performances were assessed by two raters – a near-native proficiency English teacher and a high-proficiency graduate student in the English department. As the Cronbach's alpha for the interrater reliability of the two raters was sufficiently high ($\alpha = .91$), the rest of the scoring was done by a single rater, the English teacher. Once the human rating of the participants' pronunciation was completed, the scores were converted into percentages (where 4 = 100%), and a Pearson product-moment correlation was conducted between the pronunciation scores by human raters and the average transcription accuracy scores of the STT applications to see how closely the human rater and STT applications assessed NNS English speech.

4.3.1 Correlation Between STT Application Evaluation and Human Rater Evaluation of NNS Pronunciation

To compare the STT accuracy rates with the human assessment scores on the same percentage scale, we converted the human scores (using a pronunciation rubric) into a percentage and then employed the Pearson correlation test. The overall correlation coefficient across the three speaking tasks revealed a sufficiently high correlation between STT application assessment and human rater assessment ($r = .69$). In other words, there was a similar tendency in how a human rater and ASR technology would evaluate human speech. As shown in Table 4, in particular, the strongest

Table 3 Pronunciation evaluation rubric

Score	1	2	3	4
Evaluation criteria	The accent is strong, requires a lot of effort from a listener to understand the meaning, or some parts are unintelligible. Pronunciation errors and correction of words are present.	The accent is present, but the meaning is intelligible. Few pronunciation errors and possible hesitation.	The accent is recognizable but has occasional characteristics of major varieties of English. Pronunciation is generally free of errors and lacks prosodic features.	The accent reflects the major varieties of English (native-speaker-like pronunciation). Well-paced flow.

Table 4 Correlations between five STT applications' mean accuracy scores and human rater scores (N = 30)

	Task 1: ReadS			Task 3: Retell			Task 4: QA		
Score source	Mean (%)	*SD*	*r*	Mean (%)	*SD*	*r*	Mean (%)	*SD*	*r*
5 STT app	59.04	17.08	.72	57.60	19.53	.75	64.41	13.20	.65
Human rater	72.50	23.07		71.67	22.49		73.33	23.61	

ReadS (reading short sentences), Retell (retelling the passage), and QA (answering questions)

correlation was observed in the analysis of the Retell task ($r = .75$), which indicates that the strength of the relationship (i.e., effect size) is quite large and more than half ($r^2 = .56$) of the variance scored by the human rater can be explained by the variance scored by the five STT applications. This may be partly because the accuracy scores of the STT applications for the Retell task were the most spread out (see the *SD* of the STT applications), making the score distribution more equivalent to that of human rating. More specifically, since retelling was the hardest task of the three because it required a heavy cognitive load to recall a story in detail and then retell it in the learner's own words, both the STT applications and the human rater needed to be more detail-oriented and careful when assigning a pronunciation score. This may result in a wider score distribution and reflect more on NNS pronunciation variabilities through human and STT evaluations.

Notably, although the application accuracy rates and human assessments being compared here do not take exactly the same approach to evaluate NNS pronunciation, the overall trend of pronunciation accuracy evaluation between them is very similar; that is, the mean score of the QA task is the highest, followed by the ReadS and the Retell tasks. Thus, it is safe to say that the STT applications can be relied on in assessing NNS pronunciation when they are used for low-stakes classroom pronunciation assessment.

4.3.2 Proficiency Level (CEFR) in the Context of Speech Assessment

NNS's proficiency level is another metric that has not been discussed in the context of pronunciation. As mentioned in section 4.2, the participants' English proficiency levels varied between CEFR A1 and C1.

To examine whether proficiency level can predict the success of speech assessment by STT applications and human raters, an additional two-way mixed ANOVA test was conducted on the between-factor of proficiency (four levels) and within-factor of task (Tasks 1, 3, and 4). Figure 2 represents levels of interdependency between speaking task types and individual CEFR levels based on the pronunciation scores of the human rater, and Fig. 3 represents the scores assessed by the five STT applications.

Both figures confirm relative consistency between participants' English language proficiency levels and their assigned pronunciation scores. In other words, participants' pronunciation scores increased in line with their proficiency levels, whether

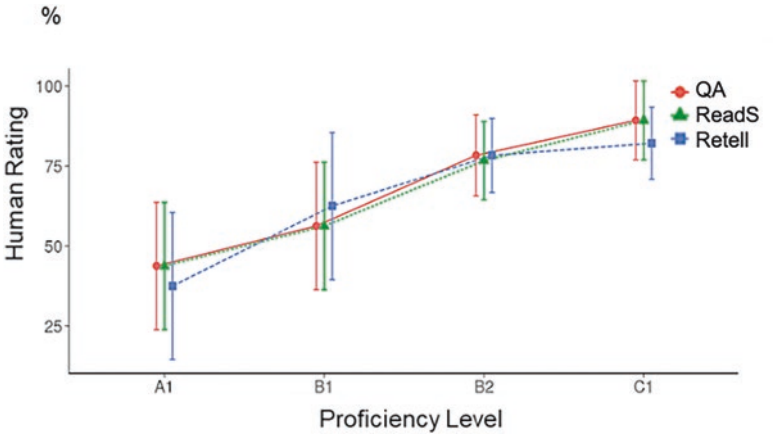

Fig. 2 Mean pronunciation score by a human rater in the context of speaking task types and proficiency rate of each participant (N = 30)

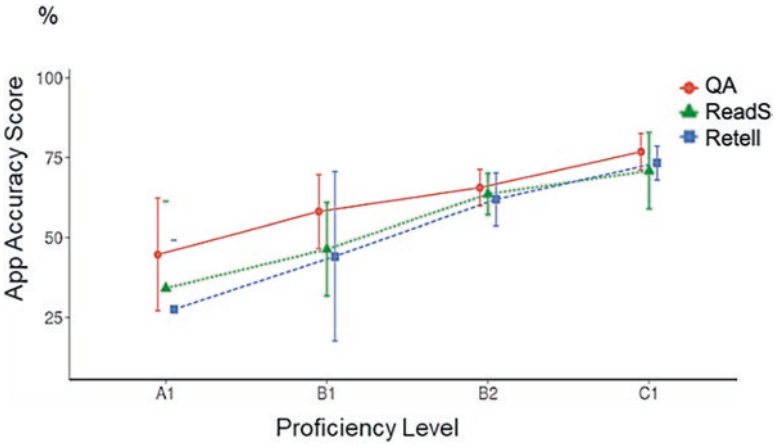

Fig. 3 Mean accuracy rate by 5 STT applications in the context of speaking task types and proficiency rate of each participant (N = 30)

they were analyzed by a human rater or an STT application. Despite the overall consistency between human and machine evaluation, the two graphs showed slight, but noticeable differences. The graph with the human rater scores shows an overall consistency of pronunciation evaluation at any language proficiency level. However, the Retell task score slightly deviates from the mean values of the other two tasks even among participants with a C1 level, which implies that the Retell task was cognitively more demanding than the other two.

Conversely, the graph with the STT application accuracy scores reports a more significant gap in the evaluation of different speaking tasks of participants with a lower English language proficiency level. In particular, participants with A1 and B1

levels of proficiency were evaluated with much less consistency than participants with a B2 level of English. At A1 and B1 levels, participants performing the QA task tended to receive higher accuracy scores than when performing the Retell task. This may be because STT applications' accuracy rates are sensitive to aspects of language production other than pronunciation, such as grammatical errors, syntactic issues, and other non-fluency features. Thus, the Retell task, which was the most difficult, affected the performance of A1-level participants more than the other tasks and the other proficiency learners.

Further analysis of the relationship between the accuracy rate of the STT applications and each task offers more insight into which of the five applications can provide the highest accuracy of evaluation, judging by the language learner's proficiency level. A Pearson correlation test between proficiency levels and each STT application showed that "Transcribe" and Windows 10's Dictation tool related most strongly to proficiency levels (see Table 5).

In particular, "Transcribe" showed the strongest correlation with proficiency in the QA task (r = .80) followed by the ReadS task (r = .74) and the Retell task (r = .73). The Windows 10's Dictation tool was also reported to have strong positive correlations, specifically in the Retell task (r = .78) and the QA task (r = .77). These results are consistent with the STT application accuracy results, where "Transcribe" and Windows 10's Dictation tool showed the best performance. In contrast, Apple's Dictation tool and Dictation.io remained consistent with lower accuracy scores; this may indicate that these applications do not have high adaptability with NNS pronunciation yet, as compared to the other applications.

4.4 Summary of Results

ASR technology, particularly STT applications, has been shown to have made significant progress in analyzing English speech by NNS. However, STT applications are affected significantly more by the English speech of NNS than by NS, despite the user's expectations (Harwell, 2018). The quality of STT analysis of NNS speech can be observed by the accuracy rate of STT transcriptions and reliability of STT assessment against human rater assessment. The five free STT applications chosen in our study recognized NNS speech at a rate of 50–70% accuracy. Of them, "Transcribe" and Windows 10's STT tools provided the best NNS speech transcription, with 68.57% and 66.96% accuracy, respectively. The other three STT

Table 5 Pearson's correlations between STT applications' accuracy scores and each task

	Google	Apple	Windows	Dictation	Transcribe
Task 1 (ReadS)	.75**	.54**	.65**	.56**	.74**
Task 3 (Retell)	.69**	.60**	.78**	.70**	.73**
Task 4 (QA)	.57**	.58**	.77**	.48**	.80**

$p < .001$**

applications—Google Docs' Voice Typing, Apple's Dictation, and Dictation.io—reported less accurate results. These accuracy rates can be affected by several factors, such as surrounding noise, participation of multiple speakers, and features of spontaneous speech. Of these, this study focused on STT applications' ability to transcribe different types of human speech. Interestingly, speaking about a comfortable and predictable topic in the QA task resulted in a higher accuracy rate. In contrast, the Retell task—another type of spontaneous speech that required more complex cognitive and memory load—resulted in more errors associated with non-fluency features. Additionally, individual pronunciation features, recognized as accented speech, also affected the accuracy rate. Transcription errors were more common in the case of speakers whose pronunciation was strongly accented due to their L1 interference. For example, Japanese speakers who had difficulty distinguishing between /l/ and /r/ encountered more transcription errors with words containing those sounds.

Regarding the reliability of the machine evaluation, it was found that human evaluation of NNS speech closely correlated with the accuracy rate of STT applications (r = .69). Despite some disparity in the assessment method and scores, the evaluation tendency remains aligned between human and machine assessment. It was also curious to see how individual proficiency levels correlated with pronunciation evaluation. A comparison between human rater evaluation and machine evaluation showed that language learners with lower proficiency levels (A1 and B1) could be more affected by inconsistent assessment of STT applications. Despite that, two applications ("Transcribe" and Windows 10's Dictation tool) still strongly correlate proficiency level and speech performance.

These results must be considered within the context of a few limitations that could have affected the study. The correlation between app accuracy scores/ human evaluation and proficiency levels of the participants (Figs. 2 and 3) is based on a rather small sample. As was mentioned above, the sample of 30 participants consisted of 4 learners of A1 level of English proficiency, 4 of B1 level, 15 of B2 level, and 7 of C1 level. The sample needs to be bigger to receive a more accurate analysis. Also, this study did not contain a control group with NS English speech, which should be addressed in future studies. Thus, the study's results analyzing NNS English speech are discussed, keeping in mind other studies on the topic.

Regardless of the limitations, the results help us recognize the current state of ASR progress in assessing NNS speech. While ASR technology still has room to grow, perhaps the absence of perfect NNS speech recognition ability could be of enormous help in recognizing pronunciation errors.

5 Practical Advice for Using STT Technology

It is important to follow some ground rules and recommendations to help English learners have the best pronunciation practice experience with ASR technology. The first thing to consider is the STT platform that will be used for pronunciation

practice. This may depend on the software and hardware available at hand. Some schools strictly control software that can be accessed on school grounds. Other institutions may be flexible but do not have funds to provide the necessary hardware to individual classes. Computer labs may be occupied with people quietly working on their projects (McCrocklin, 2015, p. 131). In this case, the teacher may ask students to use their smartphones or laptops, but they still need to ensure that all students have access to a selected STT program.

If the choice of ASR technology is not an issue, it would be a good idea to consider an STT application's accuracy and degree of strictness in terms of recognizing the speech of NNS. As observed above, tools such as Windows 10's Dictation function and "Transcribe" have a higher rate of accuracy in speech transcription, which means that students with heavier accents would receive feedback primarily on words with the least accurate pronunciation, while the rest of the speech would be recognized. Whereas if a student maintains a low rate of pronunciation inaccuracies and needs a stricter pronunciation evaluation measure, perhaps using Apple Dictation or Dictation.io would assist them in noticing a higher frequency of pronunciation errors.

Once the STT application has been selected, a teacher can work on developing appropriate tasks for pronunciation practice. These may depend substantially on the English language proficiency of the student, the task itself, and the vocabulary used in the task. If a student has a lower English proficiency level (A1-B1), the tasks need to be more controlled. For example, the tasks may include practicing pronouncing individual words (minimal pairs) or reading sentences or passages of text. Wallace (2016) suggests that teachers provide a transcript with target language that students can read to an ASR program and observe the discrepancies between the original transcript and the text transcribed by the program. From these discrepancies, a teacher can help students make conclusions about their pronunciation errors, and students can try to re-record reading the transcripts. More proficient students can attempt producing spontaneous speech, speaking into a STT application.

However, it is important to remember that the type of speaking task may noticeably affect the accuracy of ASR. As this research points out, spontaneous speech with simple utterances, such as answering easy questions, is relatively easier for ASR technology to process than spontaneous speech from memory, such as retelling. In addition, as students with a lower level of language proficiency are affected more strongly by the type of speaking task, STT applications will deliver a more significant number of transcription errors (being affected by a more considerable amount of non-fluency features and grammatical and lexical errors). Therefore, practicing pronunciation based on a preplanned text may provide the cleanest feedback regarding pronunciation analysis. In other words, reading text aloud can be a good measure of finding out about a learner's knowledge of pronunciation. However, caution is needed in that jargon or loanwords should be avoided to reduce the chance of mistranscriptions because these words are often not included in general language models used by common ASR systems (Gevirtz, 2019).

A separate comment needs to be made about the students' surroundings during pronunciation practice. As ASR technology is affected by noise or speech from

multiple speakers (Gevirtz, 2019), it is important to create a quiet environment and give students specific guidelines before practicing. These guidelines must stress the importance of keeping the environment quiet, with only one person speaking at a time. In addition, users should keep a microphone at a specified distance to avoid "breathiness," use moderate speed and volume when speaking, give shorter sentences, and reduce pause fillers such as "umm" or "ah" (Shadiev et al., 2014, p. 74). If these conditions are not kept, accuracy can be significantly reduced as it will be difficult for an STT application to recognize students' utterances.

STT applications will inevitably make some transcription errors (not only through pronunciation mistakes but also from the surrounding conditions and non-fluency features). When a particular word returns a transcription error, that would provide a good opportunity for the learner to check the pronunciation transcription of the word and listen to its correct pronunciation from a digital dictionary. McCrocklin (2015, p. 130) suggested that students should try pronouncing a word up to three times, and if it is still not recognized by the STT application, then the student should move on. Additionally, when practicing specific target words, she suggested that students focus only on the correct pronunciation of those targeted words and not pay attention to other words transcribed incorrectly. Overall, during pronunciation practice, the role of the teacher expands to providing guidance and motivation to students, as well as defining realistic objectives considering the capabilities of STT applications.

6 Conclusion

When evaluating the potential of STT applications for adult non-native learners of English to practice pronunciation, it becomes clear that ASR technology still has room to grow. The quality of an ASR's output can depend on many factors, but once outside factors are eliminated, and suitable technical conditions are met, STT applications can be excellent tools for providing feedback on a user's pronunciation. STT applications' tendency to favor the speech of NS can become a valuable measure for teachers and NNS in recognizing language learners' pronunciation inaccuracies by using the transcription function of STT applications.

The benefit of this process is multifaceted. An adult NNS, who is learning English, can receive useful feedback about their pronunciation ability by reading text to an STT application. The transcription errors visible on display can indicate mispronounced sounds and help identify ingrained pronunciation habits. For a teacher, an STT application can aid with pronunciation assessment. A teacher can prepare simple texts with target vocabulary for students with lower proficiency levels (A1-B1) or encourage higher-proficiency students to practice spontaneous speech with STT applications to help them notice their pronunciation errors. As the process of correcting human pronunciation is time-consuming and should be done on an individual basis, relying on an STT application can save time and provide feedback to a larger number of learners at the same time.

The reliability of machine speech recognition has been addressed through a research study that recognized a sufficiently high correlation between the evaluation of pronunciation by humans and STT applications. Furthermore, the assessment of the relationship between English language proficiency and STT application performance showed the potential of the STT applications to be less accurate with the NNS with lower language proficiency (A1-B1).

Therefore, teachers and language learners must wisely take advantage of the current imperfection of ASR technology until new pronunciation practice tools are developed. It is highly anticipated that the development of AI and NLP will soon result in expanding speech recognition models and increasing the accuracy rate of transcription through additional text analysis algorithms.

References

Ahn, T. Y., & Lee, S. M. (2016). User experience of a mobile speaking application with automatic speech recognition for EFL learning. *British Journal of Educational Technology, 47*(4), 778–786. https://www.researchgate.net/publication/281542912_User_experience_of_a_mobile_speaking_application_with_automatic_speech_recognition_for_EFL_learning

Altviz.co. (2019). *An introduction to automatic speech recognition* [Whitepaper]. https://bit.ly/3hVSx3b

Bajorek, J. P. (2017). L2 pronunciation in CALL: The unrealized potential of Rosetta stone, Duolingo, Babbel, and mango languages. *Issues and Trends in Educational Technology, 5*(2), 60–87. https://doi.org/10.2458/azu_itet_v5i1_bajorek

Bajorek, J. (2018). *Speech technology for language learning: Research and today's tools*. Online Language Learning Research Network (OLLReN). Cambridge University Press. https://www.researchgate.net/publication/328791102_Speech_Technology_for_Language_Learning_Research_and_Today's_Tools

Coniam, D. (1998). The use of speech recognition software as an English language oral assessment instrument: An exploratory study. *CALICO Journal, 15*(4), 7–23. https://doi.org/10.1558/cj.v15i4.7-23

Evers, K., & Chen, S. (2020). Effects of automatic speech recognition software on pronunciation for adults with different learning styles. *Journal of Educational Computing Research, 59*(4), 669–685. https://doi.org/10.1177/0735633120972011

Gevirtz, M. (2019, January 3). *The trouble with word error*. Deepgram. https://deepgram.com/blog/the-trouble-with-wer/

Godfrey, J.J., & Holliman, E. (1993). *Switchboard-1 release 2* (LDC97S62) [Data set]. Linguistic Data Consortium https://doi.org/10.35111/sw3h-rw02

Google. (n.d.). *Type with your voice.* Support.Google.Com. https://support.google.com/docs/answer/4492226?hl=en

Hachman, M. (2017, May 10). *The Windows weakness no one mentions: Speech recognition*. PC World. https://www.pcworld.com/article/3124761/the-windows-weakness-no-one-mentions-speech-recognition.html

Harwell, D. (2018, July 19). *The accent gap.* The Washington Post. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/?utm_term=.ca17667575d1

Hwang, W. Y., Shadiev, R., Kuo, T. C. T., & Chen, N. S. (2012). Effects of speech-to-text recognition application on learning performance in synchronous cyber classrooms. *Journal of Educational Technology & Society, 15*(1), 367–380. https://www.researchgate.net/publication/267263862_Effects_of_Speech-to-Text_Recognition_Application_on_Learning_Performance_in_Synchronous_Cyber_Classrooms

Ito, H. (2014). Finding practical application for speech recognition: Realizing conversations as smooth as those between native language speakers. *NII Today, 51*, 8–11. https://www.nii.ac.jp/userdata/results/pr_data/NII_Today/65_en/p8-11.pdf

Jarnow, J. (2016, April 8). *Why our crazy-smart AI still sucks at transcribing speech*. Wired. https://www.wired.com/2016/04/long-form-voice-transcription/

Kincaid, J. (2018, July 13). A brief history of ASR*: Automatic speech recognition*. Medium. https://medium.com/descript/a-brief-history-of-asr-automatic-speech-recognition-b8f338d4c0e5

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *PNAS, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Koon, L. J. (2018). Volleyball or Bareebooru? Common problems of English pronunciation for Japanese learners. *Organization for Promotion of Higher Education and Student support, 4*, 8–94. https://www.orphess.gifu-u.ac.jp/nenpou/nenpou/2018nenpo_104.pdf

Liakin, D., Cardoso, W., & Liakina, N. (2014). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal, 32*(1), 1–25. https://doi.org/10.1558/cj.v32i1.25962

McCrocklin, S. (2015). Automatic speech recognition: Making it work for your pronunciation class. In J. Levis, R. Mohammed, M. Qian & Zhou Z. *Proceedings of the 6th pronunciation in second language learning and teaching conference* (ISSN 2380-9566). Iowa State University. https://www.researchgate.net/publication/327582365

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation, 5*(1), 98–118. https://doi.org/10.1075/jslp.16034.mcc

Microsoft. (2004, February 17). Interacting with the computer using speech input and speech output. Internet Archive. https://web.archive.org/web/20040217033839/http://longhorn.msdn.microsoft.com/lhsdk/speech/speechconcepts.aspx.

O'Brien, M. G., et al. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation, 4*(2), 182–207. https://doi.org/10.1075/jslp.17001.obr

REV.com. (2020). *Speech to text report for 2020*. https://www.rev.com/blog/speech-to-text-new-research-report

Shadiev, R., Hwang, W.-Y., Chen, N.-S., & Huang, Y.-M. (2014). Review of speech-to-text recognition technology for enhancing learning. *Educational Technology & Society, 17*(4), 65–84. https://www.researchgate.net/publication/267811277

Vaughn, C., Baese-Berk, M., & Idemaru, K. (2018). Re-examining phonetic variability in native and non-native speech. *Phonetica, 76*(5), 327–358. https://doi.org/10.1159/000487269

Wallace, L. (2016). Using Google web speech as a springboard for identifying personal pronunciation problems. In J. Levis, H. Le, I. Lucic, E. Simpson, & S. Vo (Eds). *Proceedings of the 7th pronunciation in second language learning and teaching conference*, ISSN 2380-9566, Dallas, TX, October 2015 (pp. 180–186). Iowa State University.

Way, T., Kheir, R., & Bevilacqua, L. (2008). Achieving acceptable accuracy in a low-cost, assistive note-taking, speech transcription system. *Proceedings of the IASTED International Conference on Telehealth and Assistive Technologies*. ACTA Press. https://www.semanticscholar.org/paper/Achieving-acceptable-accuracy-in-a-low-cost%2C-speech-Way-Kheir/66d568ab3a8f5b95201d6ba275d2aacfdd618aef

Worthy, B. (2019, November 26). Word error rate mechanism, ASR transcription and challenges in accuracy measurement. GMR Transcription. https://bit.ly/2SrIkyU